图书情报工作
第 64 卷 第 7 期　2020 年 4 月

《左传》战争事件抽取技术研究*

■ 李章超　李忠凯　何琳
南京农业大学信息管理系　南京 210095

摘　要：[目的/意义]针对《左传》中的战争事件展开研究，对先秦历史乃至中华民族文化的研究具有重要参考价值。[方法/过程]基于框架理论构建《左传》战争事件基本框架体系，利用模式匹配法进行战争句识别，选择条件随机场模型、结合特征模板对战争时间、交战双方等 7 个命名实体进行识别和抽取，最后基于得到的结构化数据对战争事件进行分析和可视化展示。[结果/结论]研究结果表明，条件随机场模型能够较好地应用于《左传》战争事件的抽取；特征选取会影响实体识别的结果；具体内容方面，春秋时期晋国、楚国、齐国、郑国等国参战频率较高，晋国为主要进攻方，郑国为主要防守方。

关键词：《左传》　战争事件　事件抽取
分类号：G255
DOI：10.13266/j.issn.0252 - 3116.2020.07.003

1　引言

典籍，专指价值特别重大的古代汉语文本[1]，是传承中华文化的重要载体，是中华民族五千年文明的象征。数字化时代，大量典籍实现数字化并在网上公开，主要包括古籍保护数字化（原物扫描、原样复制）和古籍整理数字化两种形式[2]，其中古籍保护数字化是最主要的形式，此种形式是数字人文的基础工作，仅仅是将典籍以数字化的形式进行存储，并不利于古籍资料的检索与获取和古籍信息加工处理与深入研究。同时，已有研究的研究对象主要为现代汉语，古汉语的研究比例较小，且由于现代汉语和古汉语在词汇、句法、语法和机构等语言要素上存在明显不同，势必要对古汉语展开针对性的研究。科学研究第四范式的出现也为数字人文研究提出了新的思考：能否利用实体知识挖掘研究等新技术对典籍中的自然语言进行有效地组织，从而为历史研究提供全面、准确的典籍信息。

结合已有研究和相关文献，笔者发现对古代典籍研究具有如下特点：①研究内容方面，古汉语和历史学领域已有大量研究致力于发现典籍中实体的使用规律和构造规则，这些实体涵盖古代政治、经济、社会、军事等各个方面，如黄水清等基于文本挖掘中的条件随机场模型（Conditional Random Field，CRF），对《左传》和《国语》中的古汉语地名进行自动识别[3]。②研究方法方面，由人工方式逐渐过渡到利用计算机对典籍中的自然语言进行处理，将纸质资源转化为数字资源，并构建大型语料库供研究使用，如 C. L. Liu 等以语言模型和条件随机场模型为技术基础，挖掘 220 余部中国地方志中的传记信息[4]；钱智勇等利用隐马尔可夫模型对楚辞进行自动分词标注实验[5]。

《左传》作为中国第一部叙事型编年体史书，兼具极强的史学成就和文学价值。同时，战争是一定历史阶段国家内外部矛盾激化的典型表现[6]，是当时政治、经济、文化等要素的集中体现。因此，本文以《左传》战争事件抽取为研究主题和目标，具体包括基于框架理论的战争事件知识表示、基于规则方法的事件句抽取、实体标注研究、基于序列化标注方法的实体自动识别和战争事件可视化演示 5 个方面。

2　研究综述

2.1　事件抽取研究概况

事件抽取，是信息抽取的重要组成，以计算机技术为基础，提取自然语言文本中与某些特殊事件、事件元素或关系相关的参数，包括命名实体识别和关系抽取

* 本文系国家社会科学基金项目“基于典籍的中华传统文化知识表达体系自动构建方法”（项目编号：18BTQ063）研究成果之一。
作者简介：李章超（ORCID：0000 - 0002 - 9252 - 2142），博士研究生；李忠凯（ORCID：0000 - 0002 - 5611 - 1645），硕士研究生；何琳（ORCID：0000 - 0002 - 4207 - 3588），教授，博士，博士生导师，通讯作者，E-mail：helin@njau.edu.cn。
收稿日期：2019 - 07 - 10　修回日期：2019 - 10 - 26　本文起止页码：20 - 29　本文责任编辑：易飞

两大类别[7]。目前,事件抽取主要使用模式匹配、机器学习等方法,其中,模式匹配法是指基于模式,合理地识别、抽取某类事件,按照相应的算法匹配句子和模板,具有较高的准确率和领域专业性,如 M. Surdeanu 等开发了开放域的事件抽取系统 FSA。在机器学习法中,事件抽取的原理是通过选择和构建分类器进行分类[8],主要包括事件类型、事件元素(事件模板中的槽和事件参与者)的识别等。在现有研究中,L. C. Chieu 等将最大熵分类器引入到事件抽取中,促进了事件元素的识别[9];D. Ahn 运用 MegaM、Timbl 机器学习法对事件类型和事件元素识别进行研究,能够很好地处理 ACE 英文语料[10];赵妍妍等基于触发词集合,从文本中抽取候选事件,运用二元分类器选择合适的候选事件,并引入最大熵分类器对事件进行识别[8];于江德等利用隐马尔可夫模型对中文文本的事件抽取进行研究,基于触发词探测的方法从文本中抽取候选事件语句,根据每类事件要素的特征构建隐马尔可夫模型[11],并基于模型从语句中抽取事件要素。

此外,很多不同领域的学者结合自身研究背景对事件抽取进行了研究。比如,吴平博等关注网络事件中关键信息的抽取,利用句型模板制定信息抽取规则,基于规则确定文本中的待抽取事件,并经过时间短语识别、基本短语识别等流程抽取出质量较高的事件信息,使不同事件的分割成为可能[12];姜吉发以"知网中文词库"(HowNet)为基础,对灾害事件伤亡人员的角色信息抽取进行研究,提出一种跨语句的汉语事件信息抽取方法,并在抽取待抽取事件角色的基础上确定事件角色,得到较高的召回率和准确率[13];郑家恒等对农作物品种描述模式的获取进行研究,发现研究对象的规模和研究结果的准确性成反比[14];杨尔弘对突发事件新闻报道的信息获取进行研究,基于事件文本特征构建突发事件信息抽取模型,方便了特定的信息和信息结构的抽取[15]。

2.2 古文信息处理进展

典籍数字化迅速发展的背景下,自然语言处理技术的逐渐成熟推进了古文信息处理的发展[16]。古文信息处理是利用信息技术对古文的音、形、义进行加工,并在此基础上对古文进行深度挖掘和知识发现[5]。内容方面,古文信息资源主要通过古文数字化的形式获得,就是以数字化的形式将典籍记录、存储在计算机等可读媒介内。方法方面,随着计算机技术的不断发展,学者开始将机器学习的方法引入数字人文领域,利用其对古文进行处理和加工,主要包括古文分词和面向古文的命名实体识别。

分词是利用技术手段进行古文信息处理的基础和关键,文本分词的方法主要包括:基于规则的方法和基于统计机器学习的方法两类。基于规则的方法适用于已知句式特征等情形的结构化文本,但是对于非结构化文本的分词,很难取得理想的效果。因此,更多的学者选择基于统计机器学习的方法进行文本分词,参考基于统计机器学习对现代文本进行分词的方法,同时在常见的机器学习模型中加入词表辅助计算机进行分词,比如地名表、人名表、注疏词表等词表[17]。利用机器学习进行自动分词能够取得较好的效果,已经在很多古文本上得到应用,如《楚辞》[4]、《孟子》[18]等。

针对古汉语命名实体识别的研究同样也得到越来越多学者的关注,为古文本的知识挖掘奠定了坚实的基础。古汉语命名实体识别主要包括人名、地名等实体,其中条件随机场模型使用得最多,取得了较为理想的效果,并且已经应用在《三国演义》[19]、《春秋经传》[2]等多部典籍之中。

2.3 述评

近年来事件抽取和古文信息处理受到学者们的广泛关注,从典籍数字化发展到古文智能处理,并在自动分词、命名实体识别等方面取得一些不错的效果。目前,已有部分研究利用模式匹配和机器学习的方法对古文进行处理和分析,但在后续研究时存在针对性不强、适用性不足等弊端。因此,本文构建《左传》战争事件的基本框架体系,以模式匹配法为基础,首先构建触发词表,过滤得到候选战争句集合,再通过建立的一系列规则从候选集合中抽取出战争句,从而得到《左传》战争句语料。同时,根据之前构建的战争事件框架,基于条件随机场模型,结合《左传》文本中战争句的上下文特征、词性特征、标记特征和指示词特征,进行多次实体自动识别实验,对实验结果进行分析比较后,选取最优方案得到这些实体,具体包括:战争时间、进攻方、防守方、战争地点、战争触发原因、战争结果及援军。最后,基于以上 7 个维度,利用统计分析的方法、E-Charts 工具对数据进行分析与可视化展示。

3 研究框架

3.1 研究的总体思路

根据前述的分析和总结,本文研究框架见图 1。

首先,基于《左传》语料构建《左传》战争事件框架,对《左传》战争事件进行结构化描述、知识表示。在事件抽取阶段,本文使用效果更易控制的模式匹配

图书情报工作
第 64 卷 第 7 期　2020 年 4 月

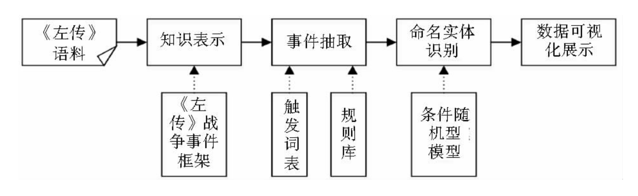


图 1　研究框架

方法进行战争句识别，通过构建触发词表并利用触发词表匹配得到初步的事件句集合，并通过规则库匹配实现事件句抽取。其次，进行命名实体的识别与抽取，通过对抽取的战争事件句集合进行观察分析，综合《左传》文本上下文窗口长度、词性和指示词等特征，基于五词位标注体系对战争事件句集合进行人工标注，并利用条件随机场模型对命名实体进行识别和抽取。

3.2　《左传》战争事件框架的建立

历史学领域认为，战争事件的主要构成要素包括战争时间、交战双方（进攻方、防守方）、战争地点、战争触发原因以及战争结果。通过对《左传》的阅读，本文发现《左传》中关于战争的描述同样涵盖以上的元素。此外，《左传》中战争事件的描述还包括对于救援事件的描述，因此本文将战争分为征战类和救援类两大类，并在救援类的战争事件中加入“援军”这个实体，能够更加具体、完整地描述《左传》中的战争事件。据此，本文构建《左传》战争事件信息基本框架，如图 2 所示：

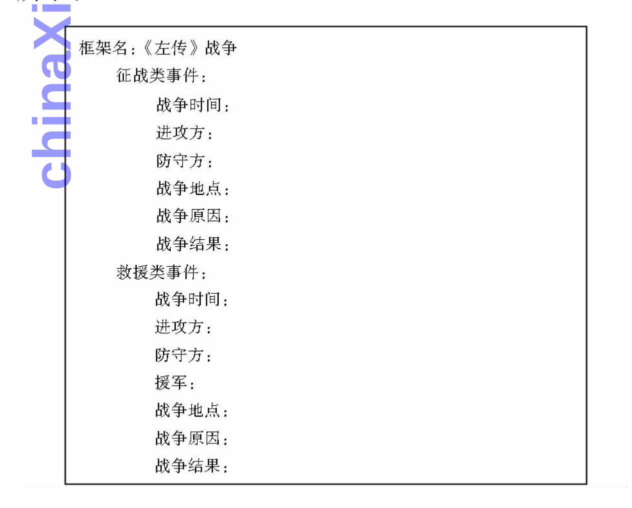


图 2　《左传》战争事件信息基本框架

示例 1：十年春，齊師伐我，戰于長勺，齊師敗績。

以此框架为基础，对示例 1 征战类事件的战争句进行抽取，得到的信息框架见图 3。

示例 2：秋，楚子圍許以救鄭，諸侯救許，乃還。

以此框架为基础，对示例 2 救援类事件的战争句进行抽取，得到的信息框架见图 4。

框架名：《左传》战争
征战类事件：
战争时间：十年春
进攻方：齊師
防守方：我
战争地点：長勺
战争结果：齊師敗績

图 3　示例 1 的抽取信息框架

框架名：《左传》战争
救援类事件：
战争时间：秋
进攻方：楚子
防守方：許
援军：諸侯
战争原因：以救鄭
战争结果：還

图 4　示例 2 的抽取信息框架

4　关键技术

4.1　战争句识别

目前，学者常用的事件句抽取方法包括：基于模式匹配的方法和基于机器学习的方法。其中，模式匹配法适用于事件句较短，且总语料数据规模较小的文本。模式匹配法以特征匹配的方法进行事件句抽取，以语言学为基础对待抽取文本进行句法分析，寻找目标主题句的规律及其与其他主题句的差异。触发词能够表达事件句之间差异的重要特征，通过构建触发词表找出包含触发词表中的句子，并在优先保证检全率的前提下，通过规则的制定，从中剔除不符合条件的句子，最终得到《左传》战争事件主题句集合。

同时，鉴于《左传》语料句法特征的复杂性，基于完备性、针对性和可行性的原则进行模式匹配，具体步骤为：①构建《左传》战争事件触发词表。《左传》中描述战争事件的文本具有一定的规律，可以根据某些与战争有关的特殊词迅速定位到待抽取的目标句，如攻、伐等，将这些特殊词进行归纳、整理，得到《左传》战争事件触发词表。②定位包含触发词的语句，抽取候选语句。根据《左传》战争事件触发词表，可以得到《左传》语料中包含特定触发词的语句，将这些语句抽取出来作为候选战争句。③剔除非战争句。最后基于制定的模式和原则，对第二步中得到的战争句和类战争句集合中的类战争句进行过滤。

4.2　触发词表构建

触发动词是进行模式抽取前预先归纳的特殊动词，是抽取战争相关要素的关键和基础。通过触发动

李章超，李忠凯，何琳.《左传》战争事件抽取技术研究[J].图书情报工作,2020,64(7):20－29.

词能够缩小文本范围,提高规则制定的效率和针对性。张秋霞对《左传》文本中征战类动词的研究,将《左传》文本中征战类动词分为9类:起兵类、交战类、攻伐类、率领类、侵扰类、戕杀类、防御类、俘获类、战果类,如表1所示[20]:

表1 《左传》征战类动词词表

动词分类	动词名
起兵类	起、兴、举、称、出、师
率领类	将、率、帅、以
交战类	合、遇、鼓、陈
侵扰类	侵、袭、犯、略、寇、掩、陵、突
攻伐类	伐、攻、军、追、逐、围、击、斗、战、要、从、伏、征
戕杀类	兵、杀、戕、戮、斩、弑
防御类	御、戍、亢、待、当、守
俘获类	取、俘、获、禽
战果类	克、胜、败、北、灭、捷、倾、崩、覆、溃、降、丧

基于对《左传》文本的阅读以及对战争句中触发词的归纳、总结,本文对张秋霞划分的9类征战类动词进行简化,剔除不在战争描述中单独使用且与其他征战类词同时使用的词,如:起、率、杀、戕等。同时,本文参考邓勇等对《左传》中战争事件的定义,加入背叛和救援两个类别[21],得到如下触发词表:"襲/v,攻/v,軍/v,戰/v,叛/v,取/v,會/v,突/v,討/v,降/v,追/v,門/v,入/v,救/v,踵/v,伐/v,敗/v,侵/v,克/v,助/v,圍/v,圍/v,滅/v,陳/v,逐/v,略/v"。完成触发词表的建立后,对整个《左传》语料进行初步辨别,从中抽取出含有触发词、且触发词词性为动词的语句,最终得到一个包含战争句和类战争句的集合。

4.3 命名实体识别

在运用条件随机场模型前期,需要一系列实验算法的设计,主要包括序列化标注、特征选择和特征模板的制定等。

4.3.1 序列化标注

基于机器学习的命名实体识别,从根本上来讲,可以转化为序列化标注的问题,也就是从语句中识别出实体,并对其中的实体进行自动标注,即对语句中的字或词语进行分类,并对其具体类别(人名、地名、时间或其他类别)进行判断。

序列化标注的单位选择是问题解决的关键,单位选择因具体任务的不同而有所区别。理论上来讲,在自动分词和词性标注的过程中,通常会选择单字为单位;在句法分析和语义角色标注的过程中,则通常会选择词语为单位。在命名实体识别的具体实践过程中,基于单字序列和基于词语序列的方法都有应用,两者也各有优劣:以单字为序列的方法,能够提供更丰富的特征,为计算机学习提供便利,但对实体边界的判断具有一定难度;以词语为序列的方法,虽然不能利用单字级别的特征,但在判断实体的边界方面具有一定的优势。另外,小语料若使用词语级的序列,会出现数据稀疏的问题,导致训练不充分,影响实体识别结果。由于《左传》语料规模较小,且目前古汉语实体识别多以单字为单位,本文选择单字作为序列化标注的单位。

以单字为单位对《左传》进行序列化标注,就是对《左传》中每个汉字进行分类。在命名实体识别的过程中,通常要在已有实体类别的基础上再进行一次分类,表示出每个汉字在实体中的位置,一般为W(单独构成实体)、B(实体首字)、M(实体中间字)、E(实体尾字)。基于此,本文定义了包含25个类别的集合Q,用于序列化标注的实体识别,如下所示:

$$Q=\left\{\begin{array}{l}\text{B-ATT, M-ATT, E-ATT, W-ATT, B-DEF, M-DEF, E-DEF,}\\ \text{W-DEF, B-TIME, M-TIME, E-TIME, W-TIME, B-HEL,}\\ \text{M-HEL, E-HEL, W-HEL, B-RES, M-RES, E-RES,}\\ \text{W-RES, B-REA, M-REA, E-REA, W-REA, O}\end{array}\right\}$$

为直观展示实体的序列化标注,本文以《左传》中的战争句"郑伯克段于鄢"为例,进行直观展示,如例(1)所示:

郑	伯	克	段	于	鄢	
X	X	X	X	X	X	例(1)

其中 $X \in Q$,命名实体识别的目的就是判断每个汉字的所属类别,例(1)中的语句经过正确的实体识别之后,结果如例(2)所示:

郑	伯	克	段	于	鄢	
B-ATT	E-ATT	O	W-DEF	O	W-LOC	例(2)

通过B、E、S等子类别中蕴含的信息得知,例(2)语句包含郑伯(人名)、段(人名)和鄢(地名)等实体。由此,通过序列化标注的方式识别出示例一个战争句中的实体。

4.3.2 特征选择

特征选择是命名实体识别的关键,能够对模型性能的发挥产生直接影响。特征,一般理解为分类模型中,能够表示类别的元素。在序列化标注模型中,汉字或者词语可以看作一种特征,另外还可以根据模型任务来增加更多特征,比如在命名实体识别过程中,可以增加姓氏、地名、词性等多项特征。本文将以古汉语典籍中的命名实体识别为任务,在汉字或词语本身特征

之外增加上下文窗口长度、标记、单词词性和实体指示词等特征。

4.3.3　特征模板制定

特征模板是条件随机场模型在训练过程中,结合所需识别的序列单位的长度,将前后字的信息及特征信息作为组合概率的信息集合。简言之,特征模板就是来定义从训练集中提取特征的方法。在著名的条件随机场开源工具 CRF + + 中,特征模板通过定义模板文件中的特征模板来提取训练文本和测试文本的特征,再通过训练集中的特征参数进行 CRF 模型计算。因此,在进行 CRF 训练时,要事先根据训练语料的特点选择合适的特征模板,进而实现模型的计算。通过不断地实验和调整,本文设定的模板窗口大小有[-1,1]、[-2,2]和[-3,3],并通过简化特征模板来观察最后的效果。

5　实验与结果分析

5.1　实验方案

5.1.1　数据来源

春秋时期战乱不断、诸侯争霸,《左传》中关于战争的记录非常全面,记叙详尽、因果完备、结构完整,具有非常大的优势。同时,《左传》文本的句法结构和形式具有规律性,处理起来比较方便;《左传》全文约 18 万余字,记录了 9 671 个词汇,具有更为丰富、且适用于系统研究的词汇量。另外,《左传》文本中的标志性词语明显,便于在文本中标注和定位,快速获取关键语句,缩短数据处理时间。为进一步保证研究的准确性,本文采用南京师范大学陈小荷团队构建的《左传》语料,已经对《左传》文本进行了校对与分词。并以此为基础,用模式匹配法识别并构建本研究所需的战争句语料库。

5.1.2　测评指标

对每次实验结果都应设定相应的评价指标或评价体系,在本实验中,我们选用的实验测评指标主要包括准确率、召回率和 F 值,具体的公式定义如下:

准确率 = 系统标注正确的属于实体的词数/系统标出的属于实体的词数 * 100%

召回率 = 系统标注正确的属于实体的词数/测试集中出现的属于实体的词数

F 值 =2 * 正确率 * 召回率/(正确率 + 召回率) * 100%

5.1.3　实验环境

(1)CRF + + 工具包的选择。目前,基于条件随机场模型的开源工具主要有 pocket crf、flexcrf 和 CRF + +。根据前人研究的经验,可知 CRF + + 是目前最受开发者欢迎的工具包,表明 CRF + + 具有较好的性能,因此,本研究选取 CRF + + 工具包作为实验工具包(CRF + +0.58 版本)。

(2)CRF + + 工具包的使用。CRF + + 工具包使用时需要用到以下 6 个文件:①crf_learn. exe:CRF + + 的训练程序;②crf_test. exe:CRF + + 的预测程序;③libcrfpp. dll:训练程序和预测程序需要使用的静态链接库;④template. data:存放特征模板的文件;⑤train. data:存放训练语料的文件;⑥test. data:存放测试预料的文件。整个 CRF + + 工具包所包含的文件如图 5 所示:

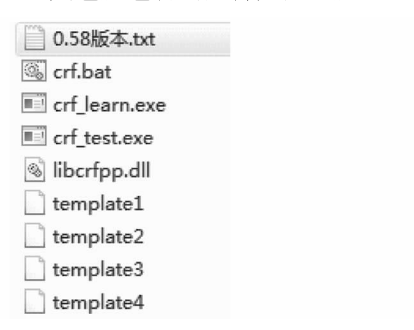


图 5　CRF + + 工具包文件构成

CRF + + 工具包对语料的格式具有非常严格的要求,一般而言,在 CRF 模型的训练文本和测试文本中包含多个 tokens(在词法分析中是标记的意思),其中每一行表示一个 token。每一行包括两列或以上的数据,第一列表示字,最后一列表示对该字的标注,中间列为可选择项(可不加,也可一个或多个),表示与该字相关的语言特征。换言之,CRF 模型的训练文本一般包括观察值、相应的特征以及状态值。加入一个特征时的训练文本示例如表 2 所示:

表 2　加入特征的训练文本示例

字符	特征	标注
四	t	B-TIME
月	t	E-TIME
,	w	O
鄭	n	B-ATT
人	n	E-ATT
侵	v	VI
衛	ns	B-DEF
牧	ns	E-DEF
,	w	O
以	c	B-REA
報	v	M-REA
東	n	M-REA
門	n	M-REA
之	u	M-REA
役	n	E-REA

表2的第一列代表字符本身(观察值),第二列为根据语料特点选择加入的特征(本示例加入的是词性特征),最后一列为该字符的字序标注(状态值)。同时,本文要用tab键(制表符)将列与列之间隔开,如果语句以标点符号结尾(如"。""?""!"等),则用换行符对句子进行空行处理,将句子与句子之间用一个空行隔开。

CRF模型的测试语料格式与训练语料大致相同,唯一的不同是测试语料可以不包含最后一列,即对第一列字符的标注列。最后通过CRF模型训练得到结果的数据格式与训练语料的格式相同,只是多了训练后得到的结果列,如表3所示:

表3 战争实体识别结果文本示例

字符	特征	标注	结果
四	t	B-TIME	B-TIME
月	t	E-TIME	E-TIME
,	w	O	O
鄭	n	B-ATT	B-ATT
人	n	E-ATT	E-ATT
侵	v	VI	VI
衞	ns	B-DEF	B-DEF
牧	ns	E-DEF	E-DEF
,	w	O	O
以	c	B-REA	B-REA
報	v	M-REA	M-REA
東	n	M-REA	M-REA
門	n	M-REA	M-REA
之	u	M-REA	M-REA
役	n	E-REA	E-REA
。	w	O	O

CRF++的执行有以下4个步骤:①将CRF++工具包自带的crf_learn.exe、crf_test.exe、libcrfpp.dll和template.data 4个文件放在同一个文件夹下,并根据需要修改特征模板文件。同时将完成格式转化的训练语料文本和测试语料文本放到前述文件夹中。②对语料进行CRF训练,并执行命令:"crf_learn template train.data model"。其中,crf_learn是条件随机场的学习算法,template是特征模板文件的文件名,train.data是训练语料文本,".data"是训练语料文本的文件格式名,model是训练过程中生成的模型文件。③利用"crf_test -m model test.data >output.txt"命令进行测试。其中,crf_test是条件随机场的测试算法,model是训练过程中生成的模型文件,test.data是测试语料文本,".data"是测试语料文本的文件格式名,output.txt是CRF测试后生成的测试结果文件。④对所生成的测试结果文件进行评估,测试命令为"conlleval.pl < output.txt"。

5.2 测评结果

5.2.1 战争句识别效果测评

本文以词性特征和标注体系符合特征为基础,采用三种不同的特征模板进行实验,其中模板一、模板二和模板三的上下文窗口长度分别为[-1,1][-2,2][-3,3]。同时在3个不同的特征模板下进行33次实验,选择每个模板实验获得的最佳效果,如表4所示:

表4 基于条件随机场的《左传》战争实体识别对照组实验结果

特征模板	正确率(Precision)	召回率(Racall)	F值(Fscore)
模板一	82.699 9%	80.484 7%	81.577 2%
模板二	80.356 8%	79.688 8%	80.021 4%
模板三	79.905 3%	79.550 1%	79.566 5%

根据表4,发现当上下文窗口长度为[-1,1]时,取得最优实体识别效果,F值达到82.699 9%,其次为上下文窗口长度为[-2,2][-3,3]时。同时,根据表4,本文发现利用条件随机场模型进行命名实体识别,不管选用哪一套特征模板,准确率和召回率均达到较高水平,均为80%左右;另外,随着上下文窗口长度的增加,正确率和召回率均呈下降趋势。

综上,本文认为窗口长度对实体识别结果具有一定影响,上下文窗口长度越长准确率越低;基于条件随机场的古汉语命名实体是可行的,并且拥有较为突出的效果。

5.2.2 不同特征模板的命名实体识别

(1)加入词性特征的CRF实体识别实验。在此次实验中加入词性特征,通过对不同特征模板进行对比实验,得到的实验结果见表5。上下文窗口长度的选择参照上述实验,其中模板一的上下文窗口长度为[-1,1],模板二的上下文窗口长度为[-2,2],模板三的窗口长度为[-3,3]。

表5 加入词性后的《左传》战争实体识别实验结果

特征模板	正确率(Precision)	召回率(Racall)	F值(Fscore)
模板一	81.348 9%	81.649 0%	81.498 6%
模板二	81.412 7%	81.407 8%	81.389 4%
模板三	81.510 2%	81.180 0%	81.344 7%

根据表5中数据,本文发现词性特征的加入使得正确率和召回率出现小幅提升,上下浮动不超过0.5%,不同窗口长度的特征模板对实体识别实验影响较小,没有出现上述实验中上下文窗口长度和准确率

呈反比的现象。究其原因,可能是词性特征对实体本身的帮助并不大,或是词性特征对实验产生一定帮助,从而造成一定的反作用。但是,通过观察,本文发现在实验中加入词性特征,使得每次实验产生的各项指标值非常接近,表明词性特征对实验效果的稳定性有一定影响。

综上,本文认为将词性特征加入实体识别,对于实验效果的影响较小。本文将在后面的实验中加入各类实体指示词,以期能够提升命名实体识别效果。

(2)加入实体指示词特征的 CRF 实体识别实验。此次实验中,只选用窗口长度为 1 和 2 的特征模板。另一方面,为了验证特征模板对实验的影响,另外编写一个窗口长度一样,但是写法有区别的模板(模板三较模板二稍做简化)进行对比实验。其中模板一的窗口长度为 1,模板二和模板三的窗口长度为 2。此次实验加入战争双方指示词、时间指示词、原因指示词、结果指示词、地点指示词和援军指示词等各类实体指示词作为特征,每个指示词为一列,出现即为 Y,否则为 N,加入标注后的示例如表 6 所示:

表 6　加入实体指示词的战争实体标注示例

字符	特征	指示词					标注
於	p	N	N	N	N	Y	O
是	r	N	N	N	N	N	O
陳	ns	N	N	N	N	N	B-REA
、	w	N	Y	N	N	N	M-REA
蔡	ns	N	N	N	N	N	M-REA
方	d	N	N	N	N	N	M-REA
睦	a	N	N	N	N	N	M-REA
於	p	N	N	N	N	Y	M-REA
衛	ns	N	N	N	N	N	E-REA
,	w	N	N	N	N	N	O
故	c	N	N	Y	Y	N	O
宋	nr	N	N	N	N	N	B-ATT
公	nr	N	N	N	N	N	E-ATT
、	w	N	Y	N	N	N	O
陳	nr	N	N	N	N	N	B-ATT
侯	nr	N	N	N	N	N	E-ATT

选取三个模板交叉实验后的最优实验结果,得到的实验结果如表 7 所示:

表 7　加入实体指示词的《左传》战争实体识别实验结果

特征模板	正确率(Precision)	召回率(Racall)	F 值(Fscore)
模板一	85.676 0%	84.321 5%	84.993 4%
模板二	87.641 9%	81.764 7%	84.601 3%
模板三	89.295 0%	80.470 6%	84.653 5%

根据表 7 的实验数据,发现在加入实体指示词作为特征之后,战争实体的识别效果出现明显提升,正确率均值达 87.54%。通过模板一和模板二的对比,发现窗口长度对实体识别效果有一定影响,窗口长度为 2 时的正确率比窗口长度为 1 时的正确率要高,但是召回率却有所下降。通过模板二和模板三的对比,发现模板写法的略微改动对实验效果也有一定影响,表明在特征模板的编写上需要进行不断的尝试和改进。同时,本文发现三个特征模板得出的 F 值很接近,综合表现相当。

5.3　数据的应用

5.3.1　战争事件统计

《左传》战争中进攻方和防守方有多种表达方法,包括以人名表示、以地名表示、以姓氏加官位进行表示等。为实现数据统一,本文通过建立人名、地名和国家进行对应的方式实现对战双方实体和国家的对应,最终得到《左传》对战双方表(示例)见表 8,通过此表中进攻方和防守方的两列数据,统计出《左传》中各国参与战争的频次,直观了解春秋时期各国参战情况。同时,本文通过 Tableau 软件生成参战国家的词云图见图 6,以此来展示各个国家参战的频率。另外,本文对《左传》中的战争事件进行统计,结果显示救援类事件共 69 件,征战类事件共 1 020 件。

表 8　《左传》对战双方表(示例)

战争句	进攻方	防守方
夏五月,鄭伯因段叛亂而伐之,克段于鄢,段出奔共	鄭	
秋八月,紀人伐夷	紀	夷
魯師敗宋師于黃	魯	宋
(冬),衛人助公孫滑伐鄭,取廩延	衛	鄭
為報衛伐鄭,(冬),鄭人以王師、虢師伐衛南鄙	衛	鄭/虢/周
夏五月,莒人入向,以薑氏還	莒	鄭
(夏),魯卿司空無駭入極,費庈父勝之	魯	極
(冬),鄭人伐衛,討公孫滑之亂	鄭	衛
夏四月,鄭祭足帥師取溫之麥	鄭	周
秋,鄭師又取成周之禾。周鄭交惡	鄭	周
春二月,莒人伐杞,取牟婁	莒	杞
夏,宋陳蔡衛伐鄭,圍其東門,五日而還(東門之役)	宋/陳/蔡/衛	鄭
秋,宋陳蔡衛及魯復伐鄭,敗鄭,取其禾而還	宋/陳/蔡/衛	鄭
(季月不詳),郕人乘衛亂之機侵衛	郕	衛
(春),曲沃莊伯以鄭人邢人伐翼,王使尹氏武氏助之,翼侯奔隨	鄭	翼
夏四月,鄭人侵衛牧,以報東門之役	鄭	衛
衛人以燕師伐鄭,鄭人以制人敗燕師於北制	衛/燕	鄭

由图 6 可知,春秋时期晋国、楚国、齐国、郑国、鲁国、卫国、宋国、吴国、秦国、陈国、蔡国等国的参战频率较高,是春秋时期战争的主要参与者。战争频率词云



图 6 《左传》战争频率词云

图总体上描述了春秋时期各国的相对参战次数，无法针对进攻方和防守方进行细分。因此，本文对各国进攻和防守的次数(单独超过 20 次)进行统计，结果见图 7 和图 8，综合二图可知晋国主要作为进攻方参与战争，主动进攻 152 次，防守 41 次；郑国主要以防守方参与战争，但其进攻次数也相对较多。

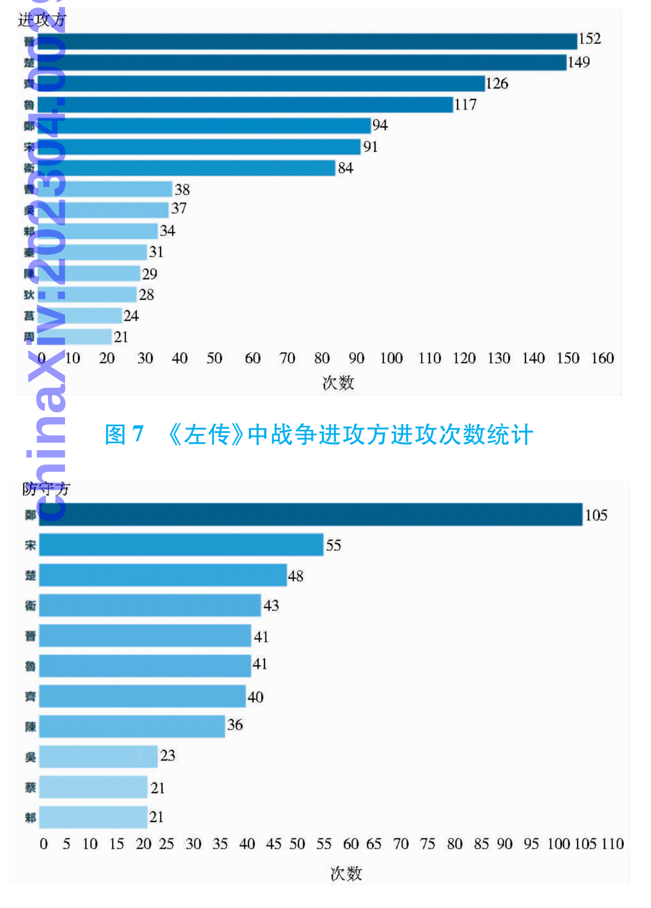


图 7 《左传》中战争进攻方进攻次数统计

图 8 《左传》中战争防守方防守次数统计

同时，本文对交战地点(发生战争超过 10 次)进行统计，见图 9。由图 9 可知，郑国、宋国和卫国作为防守方参战的次数较多，因此在他们国土上交战的次数也是最多的。剩余的作为战争地点的多是一些夹在交战双方中间的国家，如夹在郑国和宋国之间的陈国和许国。

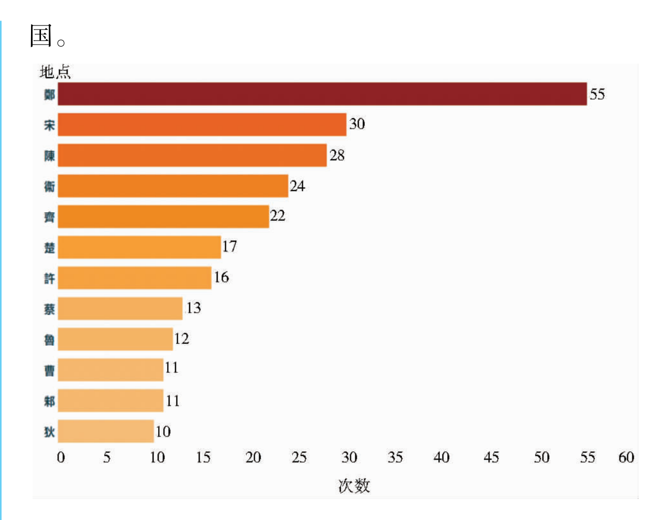


图 9 《左传》中战争地点统计

5.3.2 春秋时期地图

本文设计了一个通过 HTML、CSS 和 E-Charts3 种技术实现的《左传》战争动态展示，能更直观地了解春秋时期的大小战事，如图 10 所示：

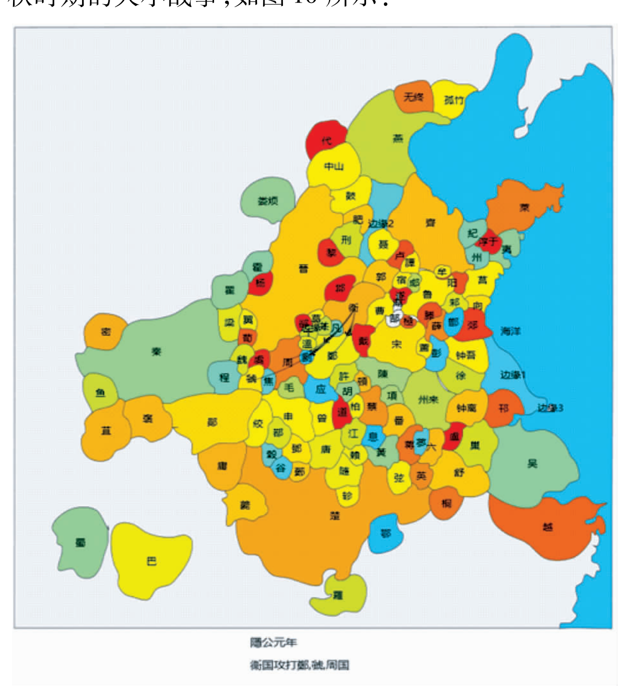


图 10 《左传》战争动态地图

《左传》战争动态地图生产步骤如下：首先，将网络中的春秋地图的图片资源转化为可以在 HTML 和 E-Charts 中使用的矢量地图，矢量化的过程中使用 Arc-GIS 作为工具，通过加载底图、新建 SHP 文件、设置各个面的属性值等步骤得到一个 SHP 格式的春秋矢量地图。其次，通过 Mapshaper 工具将 SHP 格式的地图转换为 E-Charts 能够解析的 JSON 格式的地图数据。第三，利用纯 Javascript 的图表库 E-Charts，将战争数据

转换为直观、可交互、个性化的可视化图表,具体步骤包括:①通过 echarts. init( )初始化 E-Charts 实例并放置在 div 容器中;②采用 JQuery 提供的获取 JSON 文件的语句 $ getjson( )异步加载地图数据和战争实体数据;③通过 setOption( )方法配置框架并装填数据后动态生成战争地图。

6 结论

先秦典籍内容丰富、思想活跃,凝结着先秦大家的思想与智慧。其中,《左传》是先秦时期最具代表性的史学著作之一,针对《左传》展开研究,能够为古汉语言学、考古学等诸多历史文学领域的研究提供帮助;同时,将《左传》作为实验语料,以期能够探索出有效的古汉语信息抽取方法,同时为自然语言处理领域提供参考。本文基于框架理论构建《左传》战争事件基本框架体系,利用模式匹配法进行战争句识别,选择条件随机场模型、结合特征模板对战争时间、进攻方、防守方、战争地点、战争触发原因以及战争结果 7 个命名实体进行识别和抽取,同时基于得到的结构化数据对战争事件进行分析和可视化展示。具有以下特点和优势:①将条件随机场模型和框架理论、特征模板、模式匹配法等理论方法结合起来,对于提高事件抽取的完备性、针对性和可行性具有较好的效果;②基于《左传》文本内容的特点,设计、选择相应的标注体系和特征模板,取得较好的实验效果;③通过多次实验,以验证不同窗口长度的特征模板、不同特征对于实验效果的影响,从而取得最优实验效果。最终,本研究得到以下结论:①条件随机场模型能够较好地应用于《左传》战争事件的抽取;②特征选取会影响实体识别的结果;③具体内容方面,春秋时期晋国、楚国、齐国、郑国等国的参战频率较高,其中晋国为主要进攻方,郑国为主要防守方。

参考文献:

[1] 黄水清,王东波. 古文信息处理研究的现状与趋势[J]. 图书情报工作,2017,61(12):43 -49.
[2] 施晨露. 是什么捆住了古籍数字化的手脚[EB/OL]. [2019 - 05 - 15]. https://www. jfdaily. com/news/detail? id = 53981 # top.
[3] 黄水清,王东波,何琳. 基于先秦语料库的古汉语地名自动识别模型构建研究[J]. 图书情报工作,2015,59(12):135 -140.
[4] LIU C L, HUANG C K, WANG H S, et al. Mining local gazetteers of literary Chinese with CRF and pattern based methods for biographical information in Chinese history[C]//Proceedings of the IEEE international conference on big data. Santa Clara: IEEE, 2015:1629 -1638.
[5] 钱智勇,周建忠,童国平,等. 基于 HMM 的楚辞自动分词标注研究[J]. 图书情报工作,2014,58(4):105 -110.
[6] 朱晓红. 先秦军事法思想研究[D]. 西安:西北大学,2010.
[7] 刘敏. 基于专业领域文献的信息抽取与新知识发现系统研究与应用[D]. 济南:山东大学,2018.
[8] 赵妍妍,秦兵,车万翔. 中文事件抽取技术研究[J]. 中文信息学报,2008,22(1):3 -8.
[9] HAI L C, NG H T. A maximum entropy approach to information extraction from semi-structured and free text[C]//Eighteenth national conference on artificial intelligence. San Jose: American Association for Artificial Intelligence, 2002.
[10] AHN D. The stages of event extraction[C]// Workshop on annotating & reasoning about time & events. Sydney: Association for Computational Linguistics,2006.
[11] 于江德,肖新峰,樊孝忠. 基于隐马尔可夫模型的中文文本事件信息抽取[C]//全国开放式分布与并行计算机学术会议论文集(下册). 南宁,2007.
[12] 吴平博,陈群秀,马亮. 基于时空分析的线索性事件的抽取与集成系统研究. 中文信息学报,2006,20(1):21 -28.
[13] 姜吉发. 一种跨语句汉语事件信息抽取方法[J]. 计算机工程,2005,31(2):27 -29.
[14] 郑家恒,王兴义,李飞. 信息抽取模式自动生成方法的研究[J]. 中文信息学报,2004(1):48 -54.
[15] 杨尔弘. 突发事件信息提取研究[D]. 北京:北京语言大学,2005.
[16] 高娟,刘家真. 中国大陆地区古籍数字化问题及对策[J]. 中国图书馆学报,2013(4):110 -119.
[17] 王嘉灵. 以《汉书》为例的中古汉语自动分词[D]. 南京:南京师范大学, 2014.
[18] 梁社会,陈小荷. 先秦文献《孟子》自动分词方法研究[J]. 南京师范大学文学院学报,2013(3):175 -182.
[19] 王铮. 基于 CRF 的古籍地名自动识别研究[D]. 南宁:广西民族大学, 2008.
[20] 张秋霞.《左传》征战类动词研究[D]. 长春:吉林大学,2009.
[21] 邓勇. 王霸:正义与秩序[D]. 武汉:武汉大学,2007.

作者贡献说明:
李章超:论文撰写与修改;
李忠凯:算法实现与数据分析;
何琳:确定论文选题,设计研究框架,提出论文修改建议。

李章超，李忠凯，何琳.《左传》战争事件抽取技术研究[J]. 图书情报工作,2020,64(7):20－29.

Study on the Extraction Method of War Events in *Zuo Zhuan*

Li Zhangchao　Li Zhongkai　He Lin

College of Information Science & Technology, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] This paper conducts research about the war incidents in *Zuo Zhuan*, it has important reference value for the study of pre-Qin history and Chinese culture. [**Method/process**] It constructs the basic framework system of the war incident in *Zuo Zhuan* based on the framework theory, uses the pattern matching method to identify the war sentence, selects the conditional random field model, and combines the feature template to identify and extract seven named entities, such as war time and warring parties. Finally, based on the obtained structured data, the war events are analyzed and visualized. [**Result/conclusion**] The research results show that the CRF model can be applied to the extraction of war events in *Zuo Zhuan*; the feature selection affects the results of entity recognition; about specific content, Jin, Chu, Qi, Zheng and other countries participated in the war more frequently. Jin was the main attacker. Zheng was the main defender during the Spring and Autumn Period.

Keywords: *Zuo Zhuan*　war event　event extraction